JOEL HUBER, DICK R. WITTINK, JOHN A. FIEDLER, and RICHARD MILLER*

In a large-scale national study, the authors evaluated the effectiveness of several preference elicitation techniques for predicting choices. The criteria for accuracy included both individual hit rates and a new measure, the mean absolute error predicting aggregate share using a logit choice simulator. The central finding is that hybrid models combining information from different preference elicitation tasks consistently outperform models based on one task. For example, ACA, a method that combines a self-explicated prior with relative preference measures on pairs, predicts choices better than full-profile conjoint when warmup tasks are lacking. However, there is no difference between the models if ACA's prior is combined with the full-profile information. Further, the most accurate method combines data from all three sources, suggesting that each preference elicitation technique taps a different aspect of the choice process in the validation task. Finally, full-profile conjoint is found to be significantly more accurate after rather than before, other preference elicitation tasks, implying that its performance can be improved with warmup exercises.

The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice

Preference elicitation methods reveal systematic components that underlie people's evaluations of objects. These methods can be classified into compositional and decompositional (Green and Srinivasan 1978; Pessemier et al. 1971). Compositional methods ask respondents to assess values for attribute levels, and use these values to build up preferences for attribute bundles or profiles (Huber 1974). Decompositional methods begin with overall evaluations of objects defined on multiple attributes and derive values for attribute levels from these evaluations. With either method one can predict choice from a broad range of alternatives specified by the domain of the original attributes. Once individual choice has been modeled, the prediction of choice shares in simulators has been of great value to managers as a way to estimate the impact of a change in product formulation and/or price (Bucklin and Srinivasan 1991; Green and Srinivasan 1990; Wittink and Cattin 1989).

Both decompositional and compositional methods typically take *judgments* as inputs. These judgments are often assumed to be intervally scaled measures of preferences or of the importances of attributes. What one cares most about is *choice*, the selection of one brand from a set of available alternatives. Though this distinction between judgment and choice may at first appear to involve merely substituting categorical choice for an assessment of degree of preference, numerous researchers have stressed the differences in the psychological demands of judgment and choice (Bettman and Park 1980; Huber and Klein 1991; Payne 1976, 1982; Tversky, Sattath, and Slovic 1988). In particular, choice has been shown to reflect a maze of heuristics in which decision makers seek to simplify the choice process through cutoff strategies and other noncompensatory processes (Johnson and Russo 1984; Klein and Bither 1987; Olshavsky and Granbois 1979). In contrast, repetitive evaluative judgments on objects or attributes may lead subjects to formulate relatively consistent compensatory rules that help them get through the task easily.

If the foregoing is true, any single preference elici-

*Joel Huber is Associate Professor, Fuqua School of Business, Duke University. Dick R. Wittink is Professor of Marketing and Quantitative Methods, Johnson Graduate School of Management, Cornell University. John A. Fiedler is Principal of POPULUS Inc., Boise ID. Richard Miller is President of Consumer Pulse, Inc., Birmingham, MI.

The authors thank Bill Boulding, Richard M. Johnson, and three anonymous *JMR* reviewers for their constructive input.

Journal of Marketing Research
Vol. XXX (February 1993), 105–114

tation process is likely to tap only a subset of the heterogeneous processes that underlie choice. Hence, a combination of different tasks may approximate choice better than any one method alone. To test this conjecture, we first examine the relative choice-predicting ability of three preference elicitation methods: full-profile conjoint, ACA (adaptive conjoint analysis; Johnson 1987), and ACA's self-explicated prior. We examine the hit rates of these three methods individually and suggest ways they may be improved. Then we use a logit model to merge these methods and assess the ability of the combined estimates to predict hit rates and aggregate choice shares.

Our study examines the ability of various preference elicitation techniques to predict individual choice. An alternative approach is the direct analysis of aggregated choices among experimentally controlled choice sets (Elrod, Louviere, and Davey 1992; Louviere and Woodworth 1983). However, such choice models, estimated by pooling data across individual respondents, cannot measure partworths at the level of the individual consumer. Our goal in mapping from preference elicitation measures to choice is to allow segmentation and value maximization at the individual level while closely approximating the market share predictions of choice models.

Three Preference Elicitation Methods

In defining three models—full-profile conjoint, ACA, and ACA's self-explicated prior—we do not claim that they precisely represent the way other researchers use these methods. They are designed to reflect commonly available preference elicitation tasks that may capture different aspects of choice. The self-explicated model we test is the first stage of the ACA method. Details of this method have been presented elsewhere (Green, Krieger, and Agarwal 1991; Johnson 1987, 1991). Briefly, the method begins with a rank order of levels within each attribute, either assumed to be homogeneous across respondents or derived individually. Respondents then rate the importance of the difference between the best and the worst levels of an attribute on a 4-point scale. These importance ratings are multiplied by the preference orders and rescaled so that the difference between the highest and lowest partworths of an attribute is equal to the attribute's importance. For example, a "very important" 4-level attribute has partworths of -1.5, $-.5$, .5, and 1.5, whereas an "extremely important" 3-level attribute has partworths of -2, 0, and 2. Clearly, this method is limited in its ability to adequately represent preference structures because of the truncation to four importance scores across, and the assumption of equal successive intervals within, attributes. Indeed, this self-explicated system was not designed as a stand-alone model, but as a reasonable starting point to be modified by subsequent paired judgments. We use it as a separate model to assess the degree to which the pairs add to ACA's ability to predict choice, and as an exemplar of a reasonably robust, if rudimentary, self-explicated model.

The second model we test is ACA, which combines these priors with graded paired comparisons on profiles. The system chooses attribute levels of the pairs according to a balanced, but not orthogonal, design. Because the pairs by themselves often do not have sufficient degrees of freedom to determine the partworths for individuals, ACA generates unique estimates by combining information from the self-explicated task with the paired preference judgments. The method provides two separate utility estimates, one from the self-explicated task alone and another resulting from the combination of the self-explicated data with the paired comparisons. A recent survey of commercial use of conjoint in Europe shows that from June 1986 to 1991, ACA had the highest frequency of use (Wittink et al. 1993).

The third preference elicitation method we examine, full profile, is reported to have been the most common form of conjoint used in commercial practice in the United States during the first half of the 1980s (Wittink and Cattin 1989). With this method, respondents rank or rate each profile given individually, a process that may encourage within-alternative additive processing of each attribute (Payne 1982). Contrast this task with the ACA's judgments on pairs. Paired comparisons have been shown to evoke a weighted additive difference process in which respondents progressively sum the weighted differences between alternatives on each attribute (Russo and Dosher 1983; Tversky 1969). Finally, these methods evoke different processing than the self-explicated task, which has been shown to focus inordinate attention, and thus weight, on less important attributes (Green and Srinivasan 1990; Shepard 1964; Tversky and Kahneman 1974; Weitz and Wright 1979). Clearly, the three tasks tap very different aspects of choice.

We do not have strong expectations as to which of these models will have the highest validity in predicting choice. We expect ACA to outperform its simple self-explicated prior because it combines two kinds of information. There is no consensus on whether the final ACA solution will do better or worse than the full-profile procedure. Past studies with relatively small sample sizes have not found a significant difference between the methods (Finkbeiner and Platz 1986; Green, Krieger, and Agarwal 1991; McBride and Johnson 1979). One reason to expect ACA to do better is simply that the self-explicated inputs make it more likely that the coefficients of the final solution will be in the expected direction. Srinivasan, Jain, and Malhotra (1983) show that full profile predictions can be improved by imposing constraints on the signs of partworths. ACA's use of the self-explicated information tends to do just that. This argument thus leads to the following hypothesis.

H_1: ACA predicts choice better than full-profile conjoint.

The next two hypotheses examine two factors that may moderate the accuracy of these preference elicitation methods. Looking across studies, we see that the performance of full-profile conjoint in comparison with that of a self-explicated model improves when it is second in the series (Srinivasan 1985). Three studies (Cattin, Her-

met, and Pioche 1982; Green, Goldberg, and Wiley 1982; Leigh, MacKay, and Summers 1984) show that conjoint predicts better than a self-explicated approach, whereas three other studies (Akaah and Korgaonkar 1983; Green, Krieger, and Agarwal 1991; Wright and Kriewall 1980) show the reverse. One difference between these studies is that the self-explicated preceded the conjoint task in the first series of studies, whereas the sequence was reversed in the second series. Accordingly, we hypothesize:

H_2: The advantage of ACA and its prior over full profile is greatest when full profile comes first.

A second task effect that may moderate the relative effectiveness of the methods is the number of parameters to be estimated, defined as either the number of attributes or the total number of attribute levels for which utility values are defined. Increasing the number of attributes or levels should have less impact on compositional than on decompositional models. Compositional models, such as ACA's self-explicated model, focus attention sequentially on each attribute, and thus may be less susceptible to accuracy loss from adding attributes. By contrast, the competing presence of new attributes in either pairwise or full-profile conjoint tasks may reduce the impact of the original attributes in a profile. Thus, consideration of human computational limits suggests that the predictive validity of decompositional models will be more sensitive than that of compositional ones to increasing the number of attributes. This expectation is consistent with the recommendation of Green and Srinivasan (1990) that full-profile conjoint be used when the number of attributes is about six. For 10 or more, they recommend using hybrid conjoint or ACA, both of which combine a conjoint task with a self-explicated one. Hence, we predict that:

H_3: Increasing the number of attributes has a greater impact on the predictive validity of decompositional methods than on that of self-explicated methods.

The final hypothesis is an overarching one, based on the previously discussed differences between ACA's self-explicated task, its pairwise task, and the full-profile task. Because these tasks represent different ways of eliciting preferences related to choice, a combination of methods may work better than any one individually. Related evidence from work done in forecasting suggests that combined measures tend to outperform individual ones in predicting sales (Makridakis and Winkler 1983) or new product market share (Silk and Urban 1978). Accordingly, we hypothesize:

H_4: Combinations of results from different methods predict choice better than individual ones.

METHOD

To provide within-subject control in comparing the methods, we asked each respondent to complete ACA (including self-explicated and paired judgments) and full-profile conjoint. After the first of these tasks, respondents made choices from various offered sets. These same holdout choices were replicated after the second conjoint task. The attributes and their levels, listed in Table 1, were based on a previous commercial conjoint study of refrigerators.

The experimental design consisted of the 2^4 between-subjects full factorial shown in Table 2. We manipulated four task characteristics, two of which are the focus of this article. First, we rotated task order by putting either ACA or full profile first. Second, we manipulated the number of attributes by changing the number of attributes represented in the full-profile and ACA tasks. The low number-of-attributes condition excluded four binary attributes (labeled F–I) from the nine attributes shown in Table 1.

The next two manipulations were designed to alter the weights, but not the effectiveness of the models in predicting choice. The first consisted of altering the number of levels of four test attributes. Rather than viewing all four levels shown in Table 1, half of the sample saw two

Table 1
REFRIGERATOR ATTRIBUTES

A	Brand name	General Electric, Sears/Kenmore, Whirlpool
B	Capacity	Cubic feet: 19, 20,[a] 21,[a] 22
C	Energy cost	Annual: \$70, \$80,[b] \$90,[b] \$100
D	Compressor	Extremely quiet, somewhat quiet,[b] somewhat noisy,[b] extremely noisy
E	Price	\$700, \$850,[a] \$1000,[a] \$1150
F	Design	Freezer on left (side by side), freezer on top
G	Warranty	1 year, 3 years
H	Refrigerant	Soft CFC (environmentally safe), chlorofluorohydrocarbon (harms environment)
I	Dispenser	Dispenses ice and water through the door, no door dispenser for ice or water

[a]Intermediate levels used in the design for half of the respondents.
[b]Intermediate levels used in the design for the other half of the respondents.

Table 2
EXPERIMENTAL DESIGN
(2^4 cross-subjects full factorial)

Order of tasks administered
ACA then full profile vs. full profile then ACA
Number of attributes
5 (A–E) vs. 9 (A–I)
Number of attribute levels
Price, capacity, energy cost, and compressor noise had either two or four levels. Each respondent saw two of these at four levels and two at two levels.
Attribute order within full profile

No. of Attributes	*Attribute order*		
5	A,C,D,B,E	vs.	A,B,C,D,E
9	A,C,D,I,B,G,F,H,E	vs.	A,B,G,C,D,F,H,I,E

attributes without their two intervening levels and the other half saw two other attributes without their intervening levels. Finally, we manipulated the presentation order of the attributes in the full-profile conjoint as shown Table 2. Though these two manipulations are interesting in that they may affect derived attribute weights, analysis of weight shifts is beyond the scope of our article (Wittink et al. 1992). The analyses to follow include the latter two factors as variables to ensure that they do not affect accuracy. However, because neither the main effects nor interactions involving these manipulations are statistically significant ($p < .05$), they are not discussed further.

The task, administered entirely by personal computer, began by displaying the relevant levels of all attributes. For full-profile conjoint, respondents evaluated 16 profiles on a 9-category likelihood-to-purchase scale ranging from "10% or less" through "50%" to "90% or more." For ACA's pairwise task, respondents indicated relative preferences between displayed pairs on a 9-point scale anchored at "strongly prefer A" and "strongly prefer B." The ACA system set the number of pairwise questions. The design with five attributes employed 10 pairs differing on two attributes and two pairs differing on three attributes, for a total of 12 paired comparison judgments to estimate 10 partworth parameters. The nine-attribute design added four more three-attribute pairs, so that ACA used 16 pairwise judgments to estimate 14 parameters.

For full profile, 16 judgments were used to estimate 10 partworth parameters in the low and 14 in the high number-of-attributes conditions. ACA also uses self-explicated information, so its effective number of judgments is greater than the number of pairs. Thus, the comparison of the effect of the number of attributes on the predictive validity of self explicated, ACA, and full profile is really a joint test of the psychological differences in information processing that may occur between the tasks as well as the impact of the increase in information as ACA automatically adjusts to different numbers of attributes.

The validation choice stimuli, reported in Table 3, were designed to resemble choices in the marketplace. Respondents indicated their most preferred alternative for each of two pairs and two triples. Additionally, for each triple, they also identified their least preferred alternative.[1] We deliberately structured this criterion task to be different from the previous preference elicitation procedures in several ways. First, the choice tasks required categorical responses, whereas the conjoint judgments used a 9-point continuum. Second, the choice stimuli reflected a subset of the levels shown in the conjoint tasks. As a result, respondents may have found that their favorite or desired attribute level was not available in the choice sets. Third, the choice alternatives were defined across five attributes, whereas in ACA preferences were provided for pairs defined on at most three attributes and for full profile there were either five or nine attributes. Finally, the choices were arrayed vertically rather than horizontally to differentiate the layout from the ACA task, and the order in which attribute information was provided departed from any of the orders used in the full-profile task. All of these differences were designed to make it difficult for a respondent to transfer a simplified decision process from either ACA or full profile to the choice task.

A unique aspect of our criterion task is that it was repeated, appearing at the end of the first preference elicitation task and then replicated after the second. Though we were concerned that respondents might simply remember and repeat their previous responses, the intervening conjoint task appears to have been complex enough that when asked what they noticed about the holdout choices, none of the 20 pretest respondents in-

[1]Information on the alternative least liked in a set is not included in this analysis as prediction of least liked choices has less managerial relevance than prediction of the most liked alternative. The results, however, remain substantially unchanged if these negative choices are included. See Huber et al. (1991) for detailed comparisons of ACA and full profile.

Table 3
HOLDOUT CHOICES

Choice set	*Brand*	*Annual energy cost ($)*	*Noise level*	*Capacity (cu ft)*	*Purchase price ($)*	*Sample choice share*
1	Whirlpool	100	Extremely quiet	22	1150	.66
	General Electric	70	Extremely noisy	19	700	.34
2	General Electric	70	Extremely quiet	22	1150	.71
	Sears/Kenmore	100	Extremely noisy	22	700	.29
3	Whirlpool	70	Extremely quiet	22	1150	.63
	General Electric	100	Extremely noisy	19	700	.19
	Sears/Kenmore	100	Extremely quiet	22	1150	.18
4	Whirlpool	100	Extremely quiet	22	1150	.47
	General Electric	100	Extremely noisy	19	700	.20
	Sears/Kenore	70	Extremely quiet	19	1150	.33

dicated that they had recognized the repeated questions as copies. The choice replication rate is 77% (80% for the pairs and 74% for the triples). Put differently, respondents given the identical choice sets altered their selections in about a fifth of the pairs and a quarter of the triples. The replication rate should depend on, among other factors, the closeness of the objects in overall utility. Clearly, this inconsistency in the choices limits the predictive ability of *any* preference elicitation method. It is important to note, however, that the results reported here do not depend on whether the first choice task or its replicate was used. Accordingly, the analysis is based on all of the choice data.

The sample consisted of 400 respondents interviewed at super-regional malls in 11 cities: Baltimore, Charlotte, Cincinnati, Cleveland, Colorado Springs, Denver, Detroit, Los Angeles, Milwaukee, Philadelphia, and Washington, DC. Individuals were approached and screened for being over 18 years of age and having a refrigerator in their home. Those passing the screen were given the opportunity to be part of a 30-minute computer-based survey for which they were paid $5. The computer then randomly assigned respondents to one of the 16 experimental conditions. Of the 400 who took part, all but seven completed all parts of the study and their data are included in the analysis.

Criteria for Comparing the Methods

Two criteria were used to compare the accuracy of full-profile, ACA, ACA's self-explicated prior and combinations of these methods. One was individual hit rates—the percentage of times each method correctly predicted each individual's first choice for the two replications of the four choices. The four replicated choices provided a 9-level (0–8) accuracy criterion for each respondent. This measure was treated as a continuous criterion variable in a regression analysis to determine how the experimental factors influenced model accuracy.

The second criterion was the mean absolute error in predicting aggregate choice share—the difference between the percentage of respondents predicted to choose each alternative and those who actually did so. For this aggregate measure, we had initially expected to use the same maximum utility rule used for hit rates. According to this rule, each individual is predicted to choose the alternative with the highest predicted utility. We rejected that option because the models that provided relatively superior performance on the hit rate criterion did relatively worse on the aggregate share criterion. In particular, the more accurate models for hit rates overpredicted high share and underpredicted low share alternatives. Elrod and Kumar (1989) provide an account for why this outcome may occur. The first-choice simulator assumes that the profile with the highest predicted utility will be chosen with 100% probability, something that was known to be false in our data because within-respondent choices replicated only 77% of the time. That is, if all respondents were accurately predicted to choose the same alternative, the inconsistency in choices would be responsible for imperfect predictive validity. In this sense, our very accurate models may have overpredicted choice shares for popular alternatives and underpredicted shares for unpopular ones.

What was needed, then, was a way to transform each person's intervally scaled preference score into a *probability* that reflects individual inconsistency in choice. We used a multinomial logit model using one parameter across all respondents. If V_{ik} is the intervally scaled preference value of alternative i in choice set s for individual k, then the multinomial logit model finds a weight b such that

$$P_{ikes} = \frac{e^{bV_{ik}}}{\sum_{j \in s} e^{bV_{jk}}} \quad (1)$$

is the maximum likelihood estimate of the actual choices. In effect, the parameter of this logistic function provides a transformation from the interval input scale to probabilities, adjusting for both inaccuracy in the preference scale and inconsistency in the choices. Once the coefficient has been determined, equation 1 can be used to predict each individual's choice probabilities (which then cannot be zero or one), and these predictions are aggregated to the entire sample.

RESULTS

We first examine differences between models taken individually in terms of hit rates, which are modeled as a function of the experimental conditions. We then compare the methods in terms of their ability to correctly predict aggregate choice share, using the disaggregate logit simulator just described. Finally, we examine the effectiveness of various combinations of the three core models.

The Relative Hit Rates of the Models

Table 4 gives the proportion of the choice sets correctly predicted. An appropriate way to assess relative accuracy is to take within-respondent differences between each pair of methods and test whether the mean difference across the 393 respondents differs from zero. We included the four experimental conditions and their interactions as predictors to test whether any of these conditions moderate the differences in accuracy among the models. By this test, ACA's 6-point advantage over full profile is significant ($F_{1,377} = 12.5$, $p < .001$), as is its 4-point advantage over self explicated $F_{1,377} = 17.2$, $p < .001$). The 2-point advantage of the self-explicated model over the full profile is not significant ($F_{1,377} = .70$, $p > .20$). These results must be qualified because of a significant interaction with task order in the ACA versus full profile contrast ($F_{1,377} = 11.1$, $p < .001$). There are also significant interactions for both order and number of attributes in the full profile versus self explicated contrast ($F_{1,377} = 11.2$, $p < .001$ and $F_{1,377} = 5.1$, $p = .02$).

Table 4
IMPACT OF THE NUMBER OF ATTRIBUTES AND TASK ORDER ON HIT RATES

	Proportion of first choices correctly predicted		
	Full profile method	*Self-explicated method*	*ACA method*
Average	.61	.63	.67[a]
Task order			
First	.56[b]	.62	.67
Second	.66	.64	.68
Number of attributes			
Five	.66[c]	.63	.70[c]
Nine	.57	.63	.65

[a]Difference between average ACA and the other two models significant at $p < .05$.
[b]Difference between full profile and the other two models significant at $p < .05$.
[c]Difference between five and nine attributes significant at $p < .05$.

Table 4 shows the means for these interactions. When full profile is first, its hit rate of 56% is less than either self explicated's 62% or ACA's 67%. However, when it is second its hit rate of 66% is greater than self explicated's 64%, though still slightly inferior to ACA's 68%. A simple *t*-test comparing respondents who had ACA first with those who had full profile first is significant ($t_{391} = 4.0$, $p < .001$), as is the similar contrast between full profile and self explicated ($t_{391} = 2.1$, $p < .04$). However, parallel tests comparing methods when they were second are not significant ($t_{391} = .95$ and $t_{391} = -.5$, respectively).

Finally, consider the interaction between method and number of attributes. Table 4 shows that full profile's 9-point accuracy reduction due to an increase in the number of attributes is significant ($t_{392} = 3.7$, $p < .001$), as is ACA's 5-point reduction ($t_{392} = 1.9$, $p < .05$), but the number of attributes had virtually no effect on self explicated ($t_{392} = .24$, $p > .2$). Thus, increasing the number of attributes from five to nine in our study had minimal impact on the predictive validity of the self-explicated approach and had the strongest impact on the results of full profile. As ACA is a mixture of self-explicated and pairwise conjoint judgments, its moderate loss of accuracy must come from the pairs because the accuracy of the self-explicated component is unaffected by the number of attributes.

These results generally support the four hypotheses. H_1, that ACA predicts choice better than full-profile conjoint, receives qualified support. When full profile is the first task, ACA is significantly more accurate; however, when full profile is second, the difference in favor of ACA is not significant. H_2, that the advantage of ACA and self explicated over full profile is greatest when there are no warmup tasks, predicts this pattern. Full profile is significantly more sensitive to being first than either of the other two methods. H_3, that additional attributes have a greater impact on the accuracy of decompositional than on that of compositional models, is also support. Finally, H_4, that combined methods outperform individual ones, is supported by the fact that ACA, which combines the self-explicated model and graded paired judgments, is significantly more accurate than full profile or self explicated alone. In the next section we present further evidence in favor of H_4.

Accuracy in Predicting Aggregate Choice Shares

Though ACA appears to be better at predicting individual hit rates than full profile, whether this pattern also occurs in the prediction of aggregate choice shares remains to be shown. The last column of Table 5 gives the mean absolute errors of the models, which are computed from the difference between the actual choice shares shown in Table 3 and the predicted values from the logit simulator. The mean absolute error is 5.7 for ACA, 6.2 for self explicated, and 9.8 for full profile, relative to an average share of 42. These aggregate accuracy measures thus parallel the results of model comparisons based on hit rates.

In addition to the adjustment for individual unreliability in choice, a further advantage of the logit model is that it enables us to define combinations of models that best predict hit rates and aggregate choices without having to resort to arbitrary combination strategies. This property allows an expanded test of H_4, that combined methods outperform individual strategies. We tested whether the combination of self explicated with full profile is an improvement over either alone, and whether the combination of all three methods is better than any two. Prior to the logit analyses, the predicted utilities of the profiles for the three methods were standardized across respondents to have zero mean and unit standard deviation.

The results of both tests are consistent with the general hypothesis. The combination of self explicated and full profile has significantly higher hit rates and choice share accuracy than either alone. Further, the combination of all three is best of all, with a hit rate of nearly 70% and a mean absolute error predicting share of only 3.7 percentage points.

Table 5
EFFECTIVENESS OF ALTERNATIVE MODELS PREDICTING CHOICE

Model	Criterion: Individual hit rate	Criterion: Loglikelihood of logistic model	Criterion: Mean absolute error of market shares
Chance model	.417	−2816.5	16.1
Full profile conjoint	.613[a]	−2400.7	9.8
ACA's self-explicated model	.628	−2413.4	6.2
Standard ACA	.673[a]	−2311.3	5.7
.81 full profile + .69 self-explicated	.684	−2163.3	4.4
.71 full profile + .34 self-explicated + .56 ACA	.698	−2102.3	3.7
Choice replication	.768	−1986.5	1.1

[a]Those connected by a line are not significantly different. All other pairs are significantly different at $p < .05$ level by a test of the null hypothesis that the within-respondent difference in hit rates for each pair of models is zero.

Table 5 frames these results by displaying comparable statistics for two anchor models. The worst possible model is chance—equivalent to a logistic coefficient of zero. This model results in an average hit rate of .417 (.5 for pairs, .33 for triples) and a mean absolute error of 16.1 points. In comparison with chance, all of the models do well, with strongly significant gains in the loglikelihoods.[2] At the other extreme, the replication of choice is a measure of the maximum possible performance for any model. Using choice to predict its replicate gives a hit rate of .77 with a mean absolute error in choice share of 1.1 percentage points. As one might expect, the best predictor of choice is its replicate.

DISCUSSION

Several aspects, taken together, make our study unique. First, we tested several preference elicitation methods on a large national sample of nonstudents. The product class is real and we used attributes similar to those used in a previous industry study. Second, the test is unique in that the criterion (choice among holdout profiles) was replicated within respondents. This replication permitted an assessment of the maximum predictive performance of any model. Finally, our study is one of the few in which both hit rates and choice share accuracy have been used as criteria to differentiate models.

The results of the study have implications for the selection and implementation of various preference elicitation techniques and for ways in which the various techniques can be profitably combined. Further, the study indicates the value of replicating the choice task and using choice share accuracy as an additional differentiating criterion.

Our study indicates that ACA is better than full profile in predicting choice, particularly in the absence of a substantial warmup task. However, the practical implication of this finding must be tempered by the fact that our computer implementation of full profile may not reflect normal practice. There was no deck of cards for respondents to sort, nor could respondents adjust their judgments on one profile on the basis of what they saw later. Thus, the generalizability of our results to other implementations of full profile needs to be determined. Note, however, that previous studies of conjoint have shown how robust the results are to changes in the way it is implemented (Reibstein, Bateson, and Boulding 1988).

If one chooses to use full-profile conjoint, our results provide two suggestions on how its performance can be improved. First, the sensitivity of full profile's predictive validity to the number of attributes reinforces the recommendation by Green and Srinivasan (1990) that hybrid conjoint or ACA be used when the number of attributes is much over six. Second, the order effect found indicates that full profile is likely to be more effective when it is preceded by a warmup task that familiarizes respondents with the attributes and their levels. Commercial applications of full-profile conjoint often include an initial presort of the cards. Louviere (1988), following Anderson (1975), recommends that examples of desired and undesired profiles be shown to respondents prior to the formal judgment task. Our results suggest that including one or both of these tasks should have a positive effect on the predictive effectiveness of full-profile con-

[2]For a pair of nested models, under the null hypothesis that the pair have identical predictive power, twice the change in loglikelihood is distributed as chi square with degrees of freedom equal to the difference in number of parameters of the models.

joint models, but because there have been no systematic studies comparing the impact of different warmup tasks, further research is warranted.

Though our study was designed initially to compare ACA with full-profile conjoint, its most important finding may be in expanding the methods by which different models are combined and in providing a basis for future research in that area. The combined models (ACA with self explicated; full profile with self explicated) outpredict either full profile or self explicated alone. Further, the combination of all three is significantly better than any two. An important subsequent study would compare the accuracy of a task combining a self-explicated exercise and a small number of pairwise and full-profile judgments with a task of equal duration but limited to one method. Our results suggest the former will be more accurate.

In the combination of models, we had the advantage of using the logit model to assign optimal weights to the different methods. This approach contrasts with other hybrid approaches that use the reliability of the inputs to determine relative weights (Cattin, Gelfand, and Danes 1983; Green 1984). However, we believe the primary gain comes from any combination of the different methods, with only secondary gains coming from the precise assignment of weights (see Makridakis and Winkler 1983; Moriarty 1990). As a test, we compared equal weights with the logistic weights as an alternative way to combine the three core methods. Though the equal weights model is significantly worse by the likelihood ratio test (chi square with 2 d.f. = 21.2, $p < .001$), its hit rate is very similar (.693 vs. .698), as is its average percentage error predicting choice share (4.4 vs. 3.7). Thus, though the weights determined by logistic regression are marginally better, substantial improvement is also obtained from equal weights.

Though the relative weights flowing from logit may not have been important, we found the logistic transform a strong improvement over the first-choice simulator in predicting share. This improvement may not always occur. The logistic transformation effectively adds noise to individual choice probabilities, thus making aggregate choice shares less extreme (farther from zero or one). To the extent that the partworths from a preference elicitation method already have considerable noise, adding more to it with the logistic transformation may be detrimental. In fact, with the models that have less accurate hit rates, full profile and self explicated in particular, we found that the first-choice rule does better than the "optimal" logistic transformation. However, with the more accurate models derived by combining methods, the logistic transformation consistently outperformed the first-choice rule. Accordingly, researchers are encouraged to consider the logistic transformation for market share predictions. We predict that as preference elicitation methods improve, adding error to individual choice predictions will become increasingly useful.

Traditionally, hit rates have been used in assessing the accuracy of alternative preference elicitation techniques predicting choice. In our study, the relatively small differences in hit rates between the models might have led to the conclusion that the differences, though statistically significant, are not of sufficient magnitude to be managerially important. In particular, one might have been tempted to conclude that improving a hit rate from 67% for ACA to 70% for the combined models makes relatively little difference. However, this improvement in hit rates corresponds to a 35% relative improvement in mean absolute error in choice share, from 5.7 to 3.7 points. Thus, these results indicate that a worthwhile managerial benefit can be gained by including aggregate choice share along with an individual hit rate criterion for discriminating among preference elicitation models.

The two measures are useful together because they measure very different properties of preference elicitation models. Hit rates depend primarily on the reliability of individual models, whereas choice share estimates, by aggregating over individual estimates, depend mainly on the degree to which the models provide unbiased predictions (see Hagerty 1986). It is gratifying that in our study the results are relatively invariant across the two criteria.

As in any single study, the limitations map fruitful avenues for future research. Our conclusions derive from just one durable product class, refrigerators, for which the incentive for respondents to make careful choices is unknown. An important extension would be to use product classes for which respondents face the consequences of their choices or, better yet, for which the criterion is actual purchase behavior. Further, because our tests compared only main-effect preference models, it would be useful to compare models that include attribute interaction effects. Finally, the comparisons of relative accuracy should be expanded to assess relative efficiency, the gain in accuracy relative to respondent time and effort. Multiple tasks may result in greater accuracy, but an important practical question is whether that increase in accuracy is worth the additional respondent burden.

REFERENCES

Akaah, Ishmael and Pradeep K. Korgaonkar (1983), "An Empirical Comparison of the Predictive Validity of Self-Explicated, Huber-Hybrid, Traditional Conjoint, and Hybrid Conjoint Models," *Journal of Marketing Research,* 20 (May), 187–97.

Anderson, B. F. (1975), *Cognitive Psychology: The Study of Knowing, Thinking and Learning*. New York: Academic Press, Inc.

Bettman, James R. and C. Whan Park (1980), "Effects of Prior Knowledge, Experience, and Phase of the Choice Process on the Choice Process and on Consumer Decision Processes: A Protocol Analysis," *Journal of Consumer Research,* 7 (December), 141–54.

Bucklin, Randolph E. and V. Srinivasan (1991), "Determining Interbrand Substitutability Through Survey Measurement of Consumer Preference Structures," *Journal of Marketing Research,* 28 (February), 58–71.

Cattin, Philippe, Alan Gelfand, and Jeffrey Danes (1983), "A Simple Bayesian Procedure for Estimation in a Conjoint

Model," *Journal of Marketing Research,* 20 (February), 29–35.

———, Gerard Hermet, and Alain Pioche (1982), "Alternative Hybrid Models for Conjoint Analysis: Some Empirical Results," in *Analytical Approaches to Product and Market Planning: The Second Conference*. Cambridge MA: Marketing Science Institute (October), 44–53.

——— and Dick R. Wittink (1982), "Commercial Use of Conjoint Analysis," *Journal of Marketing*, 46 (Summer), 44–53.

Elrod, Terry and S. Krishna Kumar (1989), "Bias in the First Choice Rule for Predicting Share," in *1989 Sawtooth Software Conference Proceedings*. Ketchum ID: Sawtooth Software Inc.

———, Jordan J. Louviere, and Krishnakumar S. Davey (1992), "An Empirical Comparison of Ratings-Based Conjoint Models," *Journal of Marketing Research*, 29 (August), 368-77.

Finkbeiner, Carl and Patricia J. Platz (1986), "Computerized Versus Paper and Pencil Methods," paper presented to Association for Consumer Research, Toronto.

Green, Paul E. (1984), "Hybrid Models of Conjoint Analysis: An Expository Review," *Journal of Marketing Research,* 21 (May), 155–9.

———, Stephen E. Goldberg, and James B. Wiley (1983), "A Cross Validation Test of Hybrid Conjoint Models," in *Advances in Consumer Research,* Vol. 10, R. P. Bagozzi and A. M. Tybout, eds. Ann Arbor, MI: Association for Consumer Research, 147–50.

———, Abba M. Krieger, and Manoj J. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research,* 28 (May), 215–21.

——— and V. Srinivasan (1978), "Conjoint Analysis in Consumer Behavior: Issues and Outlook," *Journal of Consumer Research*, 5 (September), 103-23.

Hagerty, Michael, R. (1986), "The Cost of Simplifying Preference Models," *Marketing Science,* 5 (Fall), 298–319.

Huber, George P. (1974), "Multi-Attribute Utility Models: A Review of Field and Field-Like Studies," *Management Science,* 20, 1339–1402.

Huber, Joel and Noreen M. Klein (1991), "Adapting Cutoffs to the Choice Environment: The Effects of Attribute Correlation and Reliability," *Journal of Consumer Research,* 18 (December), 346–57.

———, Dick R. Wittink, John A. Fiedler, and Richard L. Miller (1991), "An Empirical Comparison of ACA and Full Profile Judgments," in *1991 Sawtooth Software Conference Proceedings*. Ketchum ID: Sawtooth Software, Inc., 189–202.

Johnson, Eric and J. Edward Russo (1984), "Product Familiarity and Learning New Information," *Journal of Consumer Research,* 11 (June), 542–50.

Johnson, Richard M. (1987), "Adaptive Conjoint Analysis," in *Proceedings of the Sawtooth Software Conference on Perceptual Mapping, Conjoint Analysis and Computer Interviewing*. Ketchum, ID: Sawtooth Software, Inc., 253–65.

——— (1991), "Comment on 'Adaptive Conjoint Analysis: Some Caveats and Suggestions', " *Journal of Marketing Research,* 28 (May), 223–5.

Klein, Noreen M. and Stewart Bither (1987), "An Investigation of Utility Directed Cutoff Selection," *Journal of Consumer Research,* 14 (September), 240–56.

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

Louviere, Jordan (1988), *Analyzing Decision Making: Metric Conjoint Analysis*. Newbury Park, CA: Sage Publications, Inc.

——— and George G. Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research,* 20 (November), 350–67.

Makridakis, S. and Robert L. Winkler (1983), "Averages of Forecasts: Some Empirical Results," *Management Science,* 29 (September), 987–96.

Moriarty, Mark M. (1990), "Boundary Value Models for the Combination of Forecasts," *Journal of Marketing Research,* 28 (November), 402–17.

Olshavsky, Richard W. and Donald H. Granbois (1979), "Consumer Decision Making—Fact or Fiction," *Journal of Consumer Research,* 6 (September), 93–100.

Payne, John W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis," *Organizational Behavior and Human Performance,* 16, 366–87.

——— (1982), "Contingent Decision Behavior: A Review and Discussion of Issues," *Psychological Bulletin,* 92, 383–402.

Pessemier, E. A., P. Burger, R. D. Teach, and D. J. Tigert (1971), "Using Laboratory Brand Preference Scales to Predict Consumer Brand Purchases," *Managemment Science,* 17, 371–85.

Reibstein, David, John E. G. Bateson, and William Boulding (1988), "Conjoint Analysis Reliability, Empirical Findings," *Marketing Science,* 7 (Summer), 271–86.

Russo, J. Edward and Barbara A. Dosher (1983), "Strategies for Multiattribute Binary Choice," *Journal of Experimental Psychology: Learning, Memory and Cognition,* 9 (4), 676–96.

Shepard, Roger N. (1964), "On Subjectively Optimum Selections Among Multi-Attribute Alternatives," in *Human Judgments and Optimality,* M. W. Shelly and G. L. Bryan, eds. New York: John Wiley & Sons, Inc.

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

Srinivasan, V. (1985), "A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences," Research Paper 886, Stanford University.

———, Arun K. Jain, and Naresh K. Malhotra (1983), "Improving the Predictive Power of Conjoint Analysis by Constrained Parameter Estimation," *Journal of Marketing Research,* 20 (November), 433–8.

Tversky, Amos (1969), "Intrasitivity of Preferences," *Psychological Review,* 76, 31–48.

——— and Daniel Kahneman (1974), "Judgment Under Uncertainty, Heuristics and Biases," *Science,* 185, 1124–31.

———, Shmuel Sattath, and Paul Slovic (1988), "Contingent Weighting in Judgments and Choice," *Psychological Review,* 95 (3), 371–84.

Weitz, Barton and Peter Wright (1979), "Retrospective Self Insight on Factors Considered in Product Evaluation," *Journal of Consumer Research,* 6 (December), 280–94.

Wittink, Dick R. and Philippe Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing,* 53 (July), 91–6.

———, Joel Huber, John A. Fiedler, and Richard Miller (1993) "The Magnitude of and Explanation/Solution for the Num-

ber of Levels Effect in Conjoint Analysis," *International Journal of Research in Marketing* "Forth Coming".
———, Marco Vrienz, and Wim Burhenne (1992), "Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections," working paper, Cornell University (July).

Wright, Peter and Mary Ann Kriewall, (1980), "State of Mind Effects on the Accuracy With Which Utility Functions Predict Marketplace Choice," *Journal of Marketing Research,* 19 (August), 277–93.

Reprint No. JMR301108